碩士學位論文

Deep Learning을 活用한 學生 이탈율 방지 방안 硏究

濟州大學校 經營大學院

經營情報學科 經營情報專攻

김 종 만

2017년 8월

Deep Learning을 活用한
學生 이탈율 방지 방안 硏究

지도교수 이 동 철

김 종 만

이 論文을 經營學 碩士學位 論文으로 提出함

2017年 6月

김종만의 經營學 經營情報專攻 碩士學位 論文을
認准함

審査委員長 김 두 정 (印)

委 員 김 근 형 (印)

委 員 이 동 철 (印)

濟州大學校 經營大學院

2017年 6月

<목 차>

A Study on Prevention of Student Drop out Rate Using Deep Learning

Currently, there are many problems due to the decline in school-age population. Moreover, Korea has the largest number of universities compared to the population, and the university enrollment rate is also the highest in the world. As a result, the minimum student retention rate required for the survival of each university is becoming increasingly important.

The purpose of this study is to find out how the number of freshmen in Jeju Island from 2011 to 2014, and 8,000 students who have already selected the school, The basic direction for managing students' retention rate, which is consistently maintained from admission to graduation, is based on the data collected by gender, departure report, area of origin, grades, graduation, etc. I want to know if it is. Based on the optimal input parameters, the association analysis is performed using the apriori algorithm based on the optimal input parameters, and the most suitable training data is collected for maintenance rate management. Based on this, Deep Learning We will make the module as a basic data for development. A total of 8891 students' data were separated into training data for building a deep learning module and testing data for evaluating the model.

It is shown that the students who graduated from the specialization college and graduated from college are more likely to abandon the school in the middle of the year. These results indicate that the specialization high school is more difficult in terms of academic achievement and academic continuity, And seems to need attention.

Deep learning consisted of three hidden layers and initialized the weight by using Xavier initialization module with learning rate of 0.5, and maintained

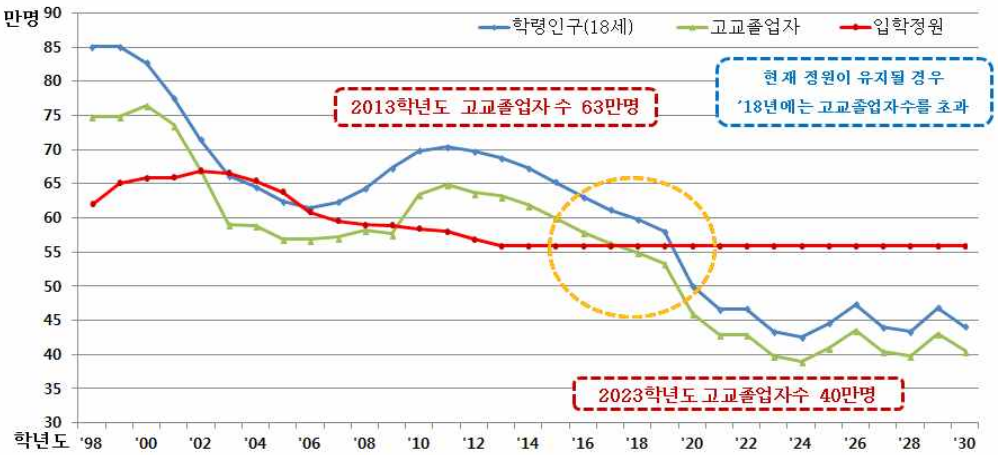
accuracy of 80%. In order to show accuracy more than 90%, it is necessary to acquire more various training data by region and university and to apply the module layer more widely.

Key Words : School Age Population Decline, Student Retention Rate, Deep Learning

Ⅰ. 서 론

현재 학령인구의 감소에 따른 많은 문제점들이 생겨나고 있다. 더욱이 우리나라는 인구대비 가장 많은 대학을 보유하고 있는 나라로서 대학 진학률 또한 세계에서 가장 높으며 OECD국가와 우리나라 대학 교육 이수율을 비교해 볼 때 OECD 평균이 42%인 반면 우리나라 이수율은 69%에 육박하고 있다(교육부, 2015년 OECD 교육지표, 2015). 또한 우리나라 공식적인 청년 실업율은 9.8%로서 역대 가장 높은 수준을 기록하고 있다. 통계청이 정의하는 공식적인 청년 실업자는 '취업을 희망하고 취업이 가능하며, 구직활동을 했음에도 실업상태인 인구를 말하며, 민간 연구기관인 현대경제연구원도 2016년 6월 체감 실업율을 34.2%로 추정하고 있다(통계청,경제활동인구조사, 2016).

<그림 1.> 학령인구 감소에 따른 고교 졸업자 수 추이



(출처 : 교육부, 고교 구조개혁 추진 계획, 2014)

위의 <그림 1>에서 보듯이 전국 기준으로 볼 때 고등학교 졸업후 진학률은 점진적으로 감소하고 있으며, 학령인구 감소에 따른 고교 졸업자 수에서도 큰 폭으로 하락하고 있는 것을 확인 할 수 있다. 2015학년도 전체 고등교육기관의 학업중단율은 7.5%로 전년 대비 0.8%p 증가한 것으로 나타난다.

일반대학의 학업 중단율은 4.1%로 전년 대비 0.2%p 증가하였고, 전문대학은 7.5%로 전년과 동일한 수준을 유지하고 있고, 방송통신대학과 사이버대학을 중심으로 한 기타 학교의 학업 중단율이 23.7%로 전년 대비 크게 증가하였다.(5.7%p↑)(교육부, 2016년 기본교육통계, 2016).

<그림 2> 전체 교육기관 학업 중단율

(출처 : 교육부, 2016년 기본교육통계, 2016)

기존의 대학생 학교이탈 관련 선행 연구는 학업 중도 이탈에 관한 연구에 많이 집중해 있는데 학업능력, 심리적요인, 전공적성, 대인관계, 거주지와 같은 특성을 분석하여 학생 개인의 특성이 대학생의 학업중단에 직접적인 영향을 미치는 요인으로 나타났다(김성식,2008; 김수연, 2007). 그리고 학생들의 가정, 대학, 대인관계, 지역사회와 같은 다양한 사회문화적 체제의 상호작용 속에서 영향을 받고 있다(김경주,송병국,박선영,2014). 이에 따라 각 대학의 생존에 필요한 최소한의 학생 유지율 관리가 점점 더 중요해 지고 있다.

본 연구에서는 위의 그림에서 보는 것과 같이 학력 인구의 점진적인 감소와 고등 교육기관에서의 학업 중단율 증가에 대비하기 위하여 신입생의 출신지역, 출신 고등학교, 성별, 나이, 학기별 성적자료를 기반으로 3 ~ 5년 정도의 학생 데이터를 수집하여 이를 기초로 연관분석을 진행하여 각 변수들 간의 연관관계를 확인 하고자 한다. 기존 연구가 설문지 중심으로 하여 주로 학생들의 심리적 요인에 대한 분석에 초점을 두었다. 학업적 능력, 심리적 특성, 진로와

적성, 대인관계, 등과 같은 개인적 특성이 학생들의 학업 이탈에 직접적인 요인으로 밝혀졌다(김성식,2008; 김수연,2007). 기존의 설문지 방식은 샘플링 과 시간적인 면에서 많은 노력이 들어가고, 또한 개인적인 특성만을 반영하는 결과를 도출하기 쉽다. 대학생들은 가정생활, 학교생활, 대인관계, 지역적 특성과 같은 다양한 사회적 문화적 체제의 상호관계 속에서 영향을 받으며 학교생활을 하고 있다. 이에 따라 최근에는 학생들의 학업 중단율을 낮추기 다양한 방법들이 시도되고 있다.

본 연구에서는 기존의 설문지 위주의 데이터 수집이 아닌, 현재 존재하는 실질적인 요인들의 데이터를 분석 하고자 한다. 예를 들면 한 대학에 입학한 학생들이 입학 당시 출신 지역 및 출신 고등학교, 성별, 나이에 따라 대학에서 휴학이나, 자퇴 또는 졸업에 대한 연관성을 분석하고, 학생 유지에 가장 영향을 미치는 변수를 찾아내어 학생 유지에 가장 적합한 변수들로 구성된 훈련 데이타를 모집하고, 이러한 데이터를 기초로 학생 유지 여부를 파악하는 Deep Learning 모듈을 개발하는데 목적이 있다. 또한 개발된 Deep Learning 모듈과 의사결정나무 모델을 활용한 이탈율 예측 정확도를 비교하여 개발된 Deep Learning 모듈이 기존의 의사결정나무 모델보다 더 좋은 정확도를 나타낼 수 있는지 알아 보았다.

구체적인 연구문제는

첫째, 기계학습의 한 분야인 Deep Learning 모듈을 활용하여 학생들이 중도탈락 없이 학기를 마칠 수 있는 확률을 측정할 수 있지 여부를 판단하고,

둘째, 측정 가능하다면 실제 학생 데이터에 의한 예측 정확도를 확인한다.

셋째, 개발된 모듈을 활용하여 통계적 접근 없이 다양한 형태의 학생정보를 활용하여 실제적인 예측 모델로 사용하여

넷째, 의사결정나무 모델과 Deep Learning 모델의 예측 정확도를 비교한다.

위의 연구문제를 해결 한다면 학생이 입학 이후에 이탈 없이 학교를 졸업할 수 있는 확률이 어느 정도인지 파악 할 수 있고, 이탈 확률이 높은 학생들을 대상으로 효과적인 지도를 할 수 있는 정보를 얻을 수 있을 것이다.

기존의 추론위주의 연구와는 달리 경험을 통해 쌓인 데이터로부터 귀납적 판단을 내는 인공지능의 학습방법을 이용하여 데이터를 기반으로 학생들이

학교를 중도에 포기할 가능성이 얼마나 되는지 판단하고, 또한 Deep Learning의 가장 큰 장점인 시간을 거듭 할수록 더욱 더 많은 경험적 데이터가 축적하여 모듈을 학습시키고, 스스로 학습하는 인공지능으로 발전한다면 더욱더 효과적인 Deep Learning 모듈로 발전해 나가길 기대한다.

Ⅱ. 이론적 배경

1. 대학생 학교이탈의 개념

대학교 학교 이탈의 개념은 1960년대 중반 미국 소규모 대학 대학생과 관련된 자퇴(withdrawal)라는 용어를 사용하는 것으로부터 시작되었다. Lembesis(1965)는 학교에 부 적응하여 자퇴할 의사가 있거나 자퇴를 하기로 한 학생들은 심리적 문제나 학교 부적응으로 규정하였다. Jeffrey(2000)는 학교이탈을 정규화된 대학교 학기에 등록하지 않은 중도탈락(dropout)으로 정의하였다. 임연욱(2007)도 학업중단을 중도탈락과 동일한 의미로 해석하고, 교육을 받을 목적으로 교육기관에 등록후 규정된 과정을 마치지 않고 그 교육기관에서 이탈하는 경우로 정의하였다. 또한 이재도(2008)은 학업이탈을 미복학, 미등록, 자퇴, 학사제적을 의미한다고 정의하였다.

본 연구에서는 대학생들이 소속하고 있는 학교에서 다음 학기로 등록을 하지 않거나 , 잠정적인 휴학을 결정할 의도가 있는 것, 즉 학교에서 학적을 상실하거나 포기하는 경우의 의도를 이탈로 규정한다.

2. 연관규칙(Association Rule)

연관 관계를 분석 할 수 있는 계산은 아이템들의 출현 빈도를 이용하여 계산하게 된다. X와 Y를 서로 공통원소가 없는 항목들의 집합이라고 하고, X->Y 를

if X then B라는 연관규칙이라고 하며, N은 전체 거래 건수, n(X), n(Y)는 항목집합 X와 Y의 거래 건수(즉, row 개수)라고 했을 때, 지지도(Support),신뢰도(Confidence), 향상도(Lift)의 정의는 아래와 같다.

2.1 지지도(Support)

두 항목 X와 Y의 지지도는 전체 거래 건수 중에서 항목집합 X와 Y를 모두 포함하는 거래 건수의 비율을 말한다.

지지도(support) s(X→Y) = X와 Y를 모두 포함하는 거래 수 / 전체 거래 수 = n(X∪Y) / N

2.2 신뢰도(Confidence)

항목집합 X를 포함하는 거래 중에서 항목집합 Y도 포함하는 거래 비율 (*조건부 확률*) 을 말한다.

신뢰도(Confidence) c(X→Y) = X와 Y를 모두 포함하는 거래 수 / X가 포함된 거래 수 = n(X∪Y) / n(X)

2.3 향상도(Lift)

항목집합 X가 주어지지 않았을 때의 항목집합 Y의 확률 대비 항목집합 X가 주어졌을 대 항목집합 Y의 확률 증가 비율을 말한다.

향상도(Lift) = 신뢰도/지지도 = c(X→Y) / s(Y)

2.4 Apriori 알고리즘

연관관계를 분석할 수 있는 알고리즘에는 모든 가능한 항목집합의 개수(M)을 줄이는 Apriori algorithm, Transaction 개수(N)를 줄여나가는 DHP algorithm, 비교하는 수(W)를 줄이는 FP-growh algorithm 등이 있다.

<표 1> 예제 데이터

ID		Items Bought
N	1	도시, 일반, 남, 졸업
	2	도시, 특수 남, 졸업
	3	도시, 검정, 여, 중퇴
	4	시골, 대졸, 여, 휴학
W		

Apriori 알고리즘은 구현이 간단하고 성능 또한 만족할 만한 수준을 보여주는 알고리즘으로 패턴 분석을 위해 자주 이용되는 알고리즘이다. 연관 관계를 계산하기 위해서 아이템들의 출현 빈도를 이용하여 계산하게 되는데 여기서는 간단히 아이템들이 동시에 출현하게 될 경우의 확률에 대해서 신뢰도 s라고 부르고, 아이템 X가 출현할 때 또 다른 아이템 Y 역시 포함되어 있을 조건부 확률을 신뢰도 c라고 할때 우선 도시 출신고, 졸업의 연관관계를 살펴보면, 도시 출신일 때 졸업할 확률은 대략 66% 정도가 나오고, 전체 데이터에서 도시와, 졸업이 일어날 확률이 50%가 된다. 따라서 지지도는 50%, 신뢰도는 66.6%가 됨을 알수 있다.

Apriori 알고리즘은 구현하기 쉽고, 이해하기 쉬우며 어느정도 만족할 만한 결과를 주기 때문에 자주 쓰이고 있으나, 이 알고리즘은 후보 집합 생성시에 아이템의 개수가 많아지면 계산 복잡도가 엄청나게 증가하게 되는 단점이 있다.

2.5 연관분석

(1) R이란?

R 프로그래밍 언어 (줄여서 R)는 통계 계산과 그래픽을 위한 프로그래밍 언어이자 소프트웨어 환경으로 뉴질랜드 오클랜드 대학의 Robert Gentleman and Ross Ihaka에 의해 개발되었으며 1997년 부터는 contributor들로 구성된 핵심 그룹에 의해서 소스코드가 관리되고 있고, 오픈GNU General Public License에 하에 배포되어 비용에 부담 없이 자유롭게 사용할 수 있는 오픈 소프트웨어이다.

R은 통계 소프트웨어 개발과 자료 분석에 널리 사용되고 있으며, 패키지 개발이 용이하여 통계 학자들 뿐만 아니라 각종 계량 연구를 하는 분야에서 널리 사

용되고 있다. 또한 R은 많은 연구자들에 의해 새롭게 만들어진 최신의 알고리즘과 로직들을 Package 형태로 제공하여 다른 어떤 통계 소프트웨어들보다도 다양한 분석방법 등을 제공한다.

(2) R의 특징[1)]

① 융통성(Versatility): R은 프로그래밍 언어이기 때문에 패키지가 미리 프로그램 해 놓은 절차에 제한 받지 않는다. 새로운 방법을 프로그램 하기가 상대적으로 쉽다.

② 대화식(Interactivity): 데이터 분석은 원래 대화식이다. 몇몇 오래된 통계패키지들은 배치 프로세싱 패러다임이 아직까지 사용법에 남아있다. R은 한 번에 하나의 프로세스를 수행한다. 분석하는 동안 보이는 것에 기초해 변경이 가능하다.

③ 호환성(Compatibility): 상업용 소프트웨어 S-Plus와 거의 호환됨.

④ 평판(Popularity): SAS가 일반영역에서 널리 통용되는 통계패키지라면 R이나 S는 통계학 연구자에게 가장 인기 있고 Finance와 Bio informatics에 특히 활용도가 높다.

(3) R의 강점

R은 자유롭게 배포할 수 있는 오픈소스로서 사용에 아무런 제한이 없다. 특히 요즘처럼 장기 불황에 있을 때 특히 많은 장점을 가지며 북미, 유럽, 호주등 대학 정규 통계수업에 사용되는 검증된 소프트웨어이다. R에는 2015년 현재 4천개가 넘는 통계분석 패키지가 있어서 최근의 분석기법이 상용 툴보다 더 빠르게 지원할 수 있으며 강력한 그래프 기능 및 데이터 처리에 있어서도 매우 강력한 툴이다.

2.6 의사결정 트리

의사결정 트리란 기계학습의 한 종류로 특정 항목에 대한 의사결정 규칙을 나무형태로 분류해 나가는 분석 기법이다. 학생들의 출신 지역 및 출신 고등학교, 성별, 나이에 따라 정상적으로 졸업할 확률을 Tree의 형태로 구분해가는 것이다. 의사결정트리의 가장 큰 장점은 분석과정이 직관적이고 이해하기 쉽다는 것이다.

1) 조완일, R의 설치 및 기본 사용법, 센소메트릭스, 2006

Deep Learning 분석의 경우 결과에 대한 설명을 이해하기 어려운 대표적인 블랙박스 모델인 반면, 의사결정트리는 분석과정을 트리형태의 구조로 쉽게 관측 할 수 있다. 그리고 수치형 및 범주형 변수를 모두 사용할 수 있다는 점, 처리 비용이 비교적 낮아 빅데이터 셋에서도 비교적 빠르게 연산이 가능하다는 점이다.

의사결정트리 분석 방법에는 통계학에 기반한(카이스퀘어, T검정, F검정) CART 및 CHAID 알고리즘이나, 기계학습 계열인 ID3, C4.5, C5.0 등의 알고리즘을 사용한다. R에는 여러 가지 의사결정트리 분석 기법(tree, rpart, party)등이 있다. tree 패키지는 binary recursive partitioning, rpart 패키지는 CART (classification and regression trees) 방법을 사용하고, party 패키지는 Unbiased recursive partitioning based on permutation tests 방법론을 사용한다.

2.7 Tensorflow

Tensorflow는 기계학급과 딥러닝을 위해 구글에서 만든 오픈소스 라이브러리이다. 텐서(Tensor)는 과학과 공학 등 다양한 분야에서 이전부터 쓰이던 개념으로서 수학에서는 임의의 기하 구조를 좌표 독립적으로 표현하기 위한 표기법으로 알려져 있지만, Tensorflow에서는 학습 데이터가 저장되는 다차원 배열을 의미한다.

Tensorflow의 특징은 다음과 같다.

- 데이터 플로우 그래프를 통한 풍부한 표현력
- 코드 수정 없이 CPU/GPU 모드로 동작
- 아이디어 테스트에서 서비스 단계까지 이용 가능
- 계산 구조와 목표 함수만 정의하면 자동으로 미분 계산을 처리
- Python/C++를 지원하며, SWIG를 통해 다양한 언어 지원 가능

(1)오퍼레이션(Operation)

그래프 상의 노드를 의미하는 오퍼레이션은 하나 이상의 *텐서*를 받을 수 있다. 오퍼레이션은 계산을 수행하고, 결과를 하나 이상의 텐서로 반환할 수 있다.

(2)텐서(Tensor)

내부적으로 모든 데이터는 텐서를 통해 표현되고, 그래프 내의 오퍼레이션 간

에는 텐서만이 전달된다.

(3)세션(Session)

그래프를 실행하기 위해서는 세션 객체가 필요한데 세션은 오퍼레이션의 실행 환경을 캡슐화한 것이다.

(4)변수(Variables)

변수는 그래프의 실행시, 패러미터를 저장하고 갱신하는데 사용된는 것으로 메모리 상에서 텐서를 저장하는 버퍼 역할을 한다.

2.8 xavier 초기화

입력값과 출력값 사이의 난수를 선택해서 입력값의 제곱근으로 나누는 초기화 방법이다.

초기화 함수의 입력값으로 초기 입력값과 출력값을 전달받아 입력값과 출력값 사이의 임의안 난수를 선택한 후 입력값의 제곱근으로 나누기하여 나온 값을 리턴하는데 여기에 균등분포로 할 것인지에 대한 옵션을 전달받아 처리한다.

```
def xavier_init(n_inputs, n_outputs, uniform=True):
Args:
n_inputs: The number of input nodes into each output.
n_outputs: The number of output nodes for each input.
uniform: If true use a uniform distribution, otherwise use a normal.
Returns: An initializer.

if uniform:
# 6 was used in the paper.
init_range = math.sqrt(6.0/(n_inputs + n_outputs))
return
tf.random_uniform_initializer(-init_range, init_range)
else:
stddev = math.sqrt(3.0/(n_inputs + n_outputs))
return
tf.truncated_normal_initializer(stddev=stddev)
```

2.9 overfitting

김성훈(2016)은 learning rate은 간단하게 보면 숫자에 불과하고 그 숫자는 프로그래밍에서처럼 "크다, 작다, 같다"의 3가지로 나누어질 수 있다.

learning rate을 너무 크지 않게 하면서도 작지 않게 조절하는 것은 무척 어렵고, 많은 경험이 필요한 영역이다. 김성훈(2016)은 overshooting이 더 안 좋은 것처럼 설명되지만, 실제로는 overshooting이 발생하면 바로 알 수 있기 때문에 문제가 되지 않는다. 오히려 learning rate이 작은 경우에 늦게 인지할 수가 있다. 마치 정상적으로 최적 값을 찾아가는 것처럼 보이니까. 두 가지 모두에 있어 여러 번의 경험을 통해 적절한지 혹은 적절하지 않은지에 대한 안목을 키우는 것이 최선이다.

적당한 learning rate을 찾는 것이 얼마나 중요한지는 여러 가지 논문에서 언급되었다. 그렇다면 learning rate을 찾기 위한 방법들에는 어떤 것들이 있을까?

김성훈(2016)에서도 여기에 대해서 몇 가지 일반론을 정리하였다.

다양한 learning rate을 사용해서 여러 번에 걸쳐 실행하면서, cost가 거꾸로 증가하는 overshooting 현상과 너무 조금씩 감소하는 현상을 확인하고 learning rate를 정하였다.

본 연구에서도 learning rate를 0.5로 정했는데 여러번의 rate조정을 거쳐 가장 최적의 결과를 나타내는 값을 정했다. 많은 연구에서 learning rate를 주관적으로 정하고, 최적의 결과를 나타내는 rate 값을 제시하였다.

overfitting을 해결하는 방법은 다음과 같다.

첫째, training data가 많을수록 overfitting에 좋다(4,000개 training data사용).

둘째, 입력으로 들어오는 변수를 개수를 줄인다(갯수 7개).

셋째, weight가 너무 큰 값들을 갖지 않도록 하는 Regularization을 사용한다.

본 연구에서 사용된 COST계산식은 다음과 같다.

$$\delta(loss) = \frac{1}{N}\sum D(s(WXi+b), Li) + \lambda \sum W2$$

W에 대해 제곱을 한 합계를 cost 함수에 더한 후 합계를 반영하는 시점에서 다양한 적용이 가능하고 λ 값을 사용해서 Panalty를 계산하였다.

Ⅲ. 선행 연구

1. 학업 중단율[2)]

권이종(2010)은 학업중단은 어떠한 이유에서든지 정규학교 교육과정을 끝내지 않고 중도에 학업을 중단하는 것을 의미한다. 학업 중단율을 줄이기 위한 노력은 그동안 중고등 학생들을 대상으로 하여 많이 연구가 되어 있으나, 고등교육기관을 상대로는 미흡한 실정이다. 그러나 최근의 교육부 통계에 따르면, 2015학년도 전체 고등교육기관의 학업 중단율은 7.5%로 전년대비 0.8% 증가하였다. 일반대학의 학업 중단율은 4.1%로 전년 대비 0.2% 증가하였고, 전문대학은 7.5%로 전년과 동일한 수준을 기록하였으나 높은 비율을 차지하고 있으며 방송통신대학과 사이버 대학을 중심으로 한 기타 학교의 학업 중단율이 23.7%로 전년 대비 크게 증가(5.7% 증가) 하여 전체 학업 중단율에 영향을 미친 것으로 파악된다.

학업중단율과 관련된 선행 연구는 대부분 고등학생들을 대상으로 많이 이루어지고 있다. 우리나라 고등학교의 정신건강이 학업성취도와 학업 중단율에 미치는 영향(방은주 외 7, 2016)에서는 학생들의 정신 건강이 학업 수행도나, 학교 이탈과 같은 교육결과와 어떤 관련성을 가지고 있는지에 대해 연구하여 낮은 학업성취도와 잦은 결석등 아동・청소년기에 보일 수 있는 정신건강 문제를 조기 발견할 수 있는 것에 대해 연구한 것이다.

또한 고등학교 학업중단율 변화의 지역별, 학교유형별 현황 및 학교 관련 요인

2) 학업중단율 = 학업중단자수 / 재적학생수 × 100

탐색(이현주 외 1, 2012)에서는 2010년까지 전국 고등학교의 학업중단율의 변화 실태를 파악하고, 이에 영향을 미치는 학교 특성 요인을 확인하는 연구를 진행하였다.

전반적으로 학업 중단율이 증가하고 있으며, 지역별로, 학교 유형별로 학업 중단율 차이에 대한 원인 분석을 하는 연구이다.

심현(2017)는 자발적으로 학업을 중단하려는 의도를 가진 국립대학의 중도 이탈자들과의 개별 면담을 통한 대학 입학에서부터 학교를 중도 탈락을 결정할때까지 경험한 심리적, 환경적 요인들을 분석하였는데 첫째 지역여건 관련 요인, 학생의 자아실현 욕구, 셋째 일과 학습의 병행하는 세가지 학업이탈 모형을 제시하였다.

이희정(2015)은 도시 학교에 제학중인 중학교 2학년부터 고등학교 1학년 사이에서 학생들이 학교 적응이 시간에 따라 어떻게 변화하는 지 학교적응을 예측하는 변수는 무엇인지를 잠재성장모형을 통하여 검증하였는데 도시지역의 학교에 다니는 청소년의 학교 적응이 증가하는 것으로 나타났고, 도시지역에서 다니는 청소년의 학교적응의 초기값을 측정하는 요인은 부모의 교육수준, 우울, 주의집중, 부모의 방임과 학대 , 수업수행 효과와 만족도, 학교폭력 가해행위와 공동체 의식 변수였고, 변화율을 예측하는 요인은 주의집중, 학업수행 효과와, 만족도, 학교폭력 가해행동과 공통체 의식 변수로 나타났다.

조영재,한석근(2015)은 부모와 학생의 배경이 학교적응에 미치는 영향을 분석하였는데 이 연구에서 학교적응 개념을 학습활동, 학교규칙, 교유관계, 교사관계로 구분하였으며 연고결과, 학교에 잘 적응하는 상위 20%집단은 학교적응에 정적 영향을 주는 요인이 높게 나타났으며, 하위 20% 집단은 부적 요인이 높은 것으로 나타났고, 부모의 방임적 양육태도는 자녀의 학교 적응과 부정적 상관관계가 있었고 학생의 학습습관 요인들은 학교 적응에 정적인 관계가 높게 나나나는 것으로 나타났다.

이병환,강대구(2014)에서는 학생들의 학교생활 부적을 요인을 파악하기 위하여 학교 부적응 학생들을 대상으로 평균과 표준편차와 위계적 회귀분석을 통하여 결과를 분석하였다. 수업에 적응하지 못하는 행동은 첫째, 부적응 행동 경험과 교사의 기대, 학업성적과 부모자녀 관계 요인이 영향을 미치는 것으로 나타났다.

둘째, 가정폭력 셋째, 학교 생활에 흥미가 떨어지는 학생일수록 학교생활에 어려움을 격고 있는 것으로 나타났다. 위와 같이 중 고등학생들을 대상으로 학업 중단율에 영향을 미치는 연구는 비교적 많이 진행 되어져 왔으나, 고등학교 이상 즉 고등교육기관을 대상으로 학업 중단율에 영향을 미치는 요인을 분석하는 연구에서는 많이 부족한 것으로 파악 되었다.

또한 학업 중단율의 영향을 연구하는 방법도 대부분 심리적 요인등을 대상으로 연구 하여 왔으며, 실제적 데이터로 접근한 연구는 없는 것으로 판단 되었다. <표2>에서 보듯이 관련 학업 성취도 및 중도 탈락관련 기존 연구에서 권이종(2010), 이희정(2015) 등 청소년 관련 연구가 많은 부분을 차지하고 있으며, 김수현(2006), 심현(20170등 고등교육기관 관련 연구는 상대적으로 적은 비율을 차지하고 있다.

본 연구에서는 기존 연구에서의 심리적인 요인에서 중도 탈락 관련 원인을 찾는 형태가 아닌 대학교 학생들의 실제적인 데이터를 활용하여, 심리적 설문조사나, 사전 조사 없이 현재의 실질적인 데이터 만을 활용하여 학생들의 이탈율을 예측하는데 활용 가능한 모듈을 개발하였다.

<표 2> 학업성취도 및 학교 중도 탈락 관련 기존 연구

구분	연구자	년도	연구주제	비고
청소년	권이종	2010	학업중단 아동과 청소년의 문제점과 해결방안	
	이병환,강대구	2014	중고등학교 학생들의 학교부적응 행동요인 분석	
	이희정	2015	도시 청소년 학교적응의 변화와 개인,가족,학교 및 지역사회 관련 예측변인	
	이지현	2015	학교 부적응에 영향을 미치는 학생 개인 및 학교 수준 요인 분석	
	정문주,김혜경,문윤희	2015	학습자가 인식한 교수자의 수업방식이 학업성취 향상요인 및 학업성취도에 미치는 영향	
	조영재,한석근	2015	부모 및 학생 배경과 학교적을 관계 분석	

고등교육기관	김수연	2006	대학생의 학업지속과 중도탈락 요인 분석	
	김성식	2008	대학생들의 학업중단 및 학교이동에 대한 탐색적 분석	
	김 경 주 , 송 병 국, 박선영	2014	지방사립대학을 중심으로 한 대학생의 학교이탈 관련 변인 탐색	
	심현	2017	국립대학 학생의 중도탈락 요인에 관한 근거이론 기반 분석	

Ⅳ. 연구 설계

1. 데이터 구성

Training data 및 Testing Data의 구성은 출신학교 , 성별, 연령, 출신지역, 통학시간, 성적등으로 구분하고, 졸업여부를 판정하기 위한 데이터로 구성하였다.

<표 3> 데이터의 구성

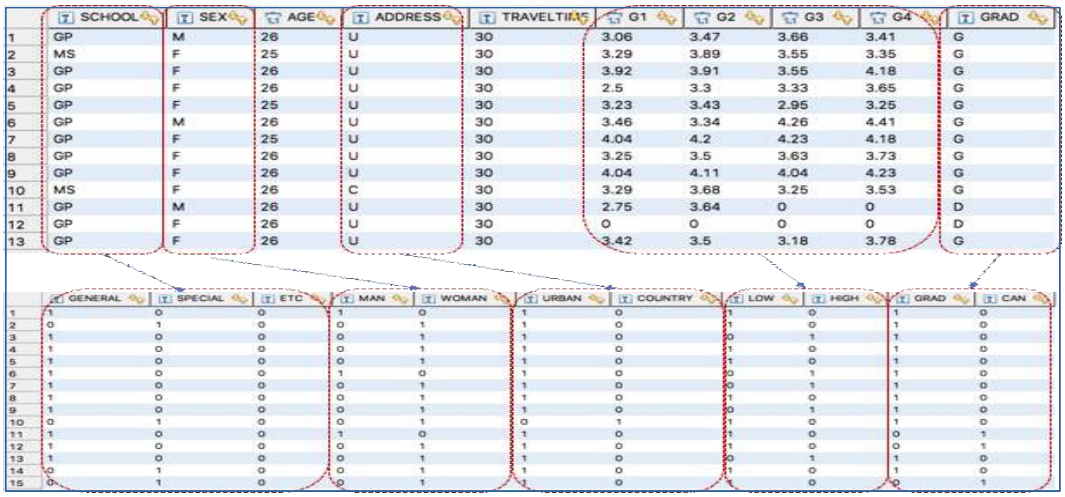
순서	항목	설명
1	school	출신학교 (binary: "GP" - 일반고 or "MS" - 특수 or “SF” - 검정 or “ET” - 기타)
2	sex	student's sex (binary: "F" - 여자 or "M" - 남자)
3	age	student's age (numeric: 15 ~ 22)
4	address	출신지역 (binary: "U" - 도시 or "C" - 시골)
5	traveltime	통학 시간 (numeric: 1 - <15 분., 2 - 15 to 30 분., 3 - 30 to 1 시간, or 4 - >1 시간이상)
6	G1	first period grade (numeric: from 0 to 20)
7	G2	second period grade (numeric: from 0 to 20)
8	G3	final grade (numeric: from 0 to 20, output target)
9	grad	졸업 여부 (binary: “G” - 졸업 or “H” - 휴학 or “D” - 중퇴(제적) or “C” - 수료 or “P” - 재학

출신학교는 일반고, 특성화고, 검정고시 그리고 대졸 및 외국인 학교로 구분하였으며 출신지역은 도시출신과 시골지역으로 구분하였고, 통학시간은 시간대별로 범위로 구분하였으며, 졸업여부는 정상졸업과 자퇴,장기 휴학상태,제적인 경우로 구분하였다<표 3>.

2. 데이터 특성

2011학년도부터 2014 학년도 까지의 제주시내 대학교 신입생 및 재학생 대상 8,000명을 대상으로 하여 이름, 학번 등의 개인정보 데이터를 제외하고, 학생들의 출신 지역, 출신 고등학교, 성별, 나이, 주소, 성적등의 데이터를 활용하였다. 이 데이터는 학생이 대학 입학 시 기록한 것으로 성별, 나이, 출신고 등을 나타내는 변수들과 통학거리를 나타내는 변수들, 그리고 학기별 성적 데이터를 나타내는 변수로 구성되어 있으며 학생이 정상적으로 학교를 졸업 했을 경우 와 그렇지 않을 경우를 고려하여 데이터를 구성하였다. <표 4>에서와 같은 구조로 구성된 범주형 데이터를 이용하여 연관분석을 실행하기 위하여 이항데이터(binary data)로 변환하여 사용하였다. 출신학교 여부, 성별, 출신지역여부, 졸업여부 등은 범주형 데이터 이며 연령(age)는 연속형 데이터(continuous data)에 속하는데 이를 '1','0'의 두 개의 값만 가지는 변수로 변환하는 이항 변수화는 다음과 같다.

<표 4> 범주형/연속형 데이터의 이항 변수화(Binarization)



	SCHOOL	SEX	AGE	ADDRESS	TRAVELTIME	G1	G2	G3	G4	GRAD
1	GP	M	26	U	30	3.06	3.47	3.66	3.41	G
2	MS	F	25	U	30	3.29	3.89	3.55	3.35	G
3	GP	F	26	U	30	3.92	3.91	3.55	4.18	G
4	GP	F	26	U	30	2.5	3.3	3.33	3.65	G
5	GP	F	25	U	30	3.23	3.43	2.95	3.25	G
6	GP	M	26	U	30	3.46	3.34	4.26	4.41	G
7	GP	F	25	U	30	4.04	4.2	4.23	4.18	G
8	GP	F	26	U	30	3.25	3.5	3.63	3.73	G
9	GP	F	26	U	30	4.04	4.11	4.04	4.23	G
10	MS	F	26	C	30	3.29	3.68	3.25	3.53	G
11	GP	M	26	U	30	2.75	3.64	0	0	D
12	GP	F	26	U	30	0	0	0	0	D
13	GP	F	26	U	30	3.42	3.5	3.18	3.78	G

	GENERAL	SPECIAL	ETC	MAN	WOMAN	URBAN	COUNTRY	LOW	HIGH	GRAD	CAN
1	1	0	0	1	0	1	0	1	0	1	0
2	0	1	0	0	1	1	0	1	0	1	0
3	1	0	0	0	1	1	0	0	1	1	0
4	1	0	0	0	1	1	0	1	0	1	0
5	1	0	0	0	1	1	0	1	0	1	0
6	1	0	0	1	0	1	0	0	1	1	0
7	1	0	0	0	1	1	0	0	1	1	0
8	1	0	0	0	1	1	0	1	0	1	0
9	1	0	0	0	1	1	0	0	1	1	0
10	0	1	0	0	1	0	1	1	0	1	0
11	1	0	0	1	0	1	0	1	0	0	1
12	1	0	0	0	1	1	0	1	0	0	1
13	1	0	0	0	1	1	0	0	1	1	0
14	0	1	0	0	1	1	0	1	0	1	0
15	0	1	0	0	1	1	0	1	0	0	1

본 연구에 사용된 데이터의 특성은 <표 5>에서 보는 바와 같이 성별로는 남자가 52%, 여자가 48%이며 출신학교는 일반 고등학교가 67% 특성화고가 24% 검정고시등 기타 고등학교가 8.4%의 비율로 나타났다. 정상졸업인 경우 62.6% 이며 제적, 중퇴, 장기휴학등 졸업하지 않은 비율이 37.4%로 나타났으며, 이중 남성인 경우 정상 졸업이 44.7%, 중도탈락이 55.3%, 여성인 경우 졸업이 79%, 중도탈락이 21%로서 남성 보다는 여성이 중도탈락 없이 학교를 정상적으로 마치는 비율이 높았다.

<표 5> 데이터의 주요 특성

주요 특성		빈도	비고
성별	남자	52%	
	여자	48%	
출신고등학교	일반고	67%	
	특성화고	24%	
	기타(검정,외국)	8.4%	
졸업여부	졸업	62.6%	남자:44.7%, 여자:79.0%
	중퇴,	37.4%	남자:55.3%, 여자:21.0%

<표 6>은 R을 활용하여 범주형으로 구성된 데이터를 이산형으로 변경하는 일련의 작업을 나타내고 있다.

출신학교, 성별, 지역, 연령 등의 구분자를 구성하여 각각의 변수에 대입한 후 연속형 값을 이산화하기위해 정렬한후 R의 data frame 함수를 사용하여 이산화형으로 나누기 위한 분할점을 구성한 후 이렇게 나뉘어진 데이터를 거래데이터(이산화형)로 변환한다.

- 연속된 변수의 값을 이산화하기 위해 정렬한다.
- 범주형을 이산화형으로 나누기 위한 분할점을 구성한다.
- 분할점으로 구성된 데이터를 이산화형(Transaction Data)로 변환한다.

<표 6> 범주형 데이터의 이산형화

```
범주형 dataset 확보 연속형 변수의 이산형화(discretization)
## categorical data -> binarization -> association rule analysis
student[["school"]]<-  c("GP","MS","SF","ET")
student[["sex"]]<-  c("Male","Female")
student[["address"]]<-  c("Unban","Country")
student[["g1"]]<-  c("low","high")
student[["grad"]]<-  c("graduate","drop")
student_mart <- as.data.frame(student)
sapply(student_mart,class)
str(student_mart)
# dataset for associtionrule : (2) transaction data format,
(거래데이터 형식으로 데이터 변환하기)
student_mart_tr <- as(student_mart,"transactions")
```

3. 데이터 분석

3.1 연관 분석

본 연구는 이론적 배경을 기반으로 실제적으로 데이터에서 포함하는 변수들이 실제 학생 이탈에 영향을 미치는지에 대한 연관성을 분석하고, 주요 연구변수들이 실제 이탈율에 영향을 미친다면 이것을 바탕으로 좀더 많은 데이터를 수집하여 실제 학생 개개인이 입학과 동시에 학기일정을 정상적으로 마칠 수 있는 확률값을 확보하기 위한 것이다.

학생들의 출신학교는 일반고등학교 출신일 경우 'GP', 취업을 목적으로 하는 특성화고 등은 'MS', 그리고 검정고시, 대졸출신, 외국고 출신 등은 'ET' 로 표시하였다. 성별은 남성일 경우 'M', 여성일 경우 'F'로 구분 하였으며 주소는 직할시를 포함한 시 지역일 경우 'U' 그 외 지역 출신일 경우 'C'로 표기 하였으며 성적은 평점이 3.3 이상인 경우 'high' 평점 3.3 미만인 경우 'low' 로 표기하였으며, 졸업여부는 정상적으로 졸업한 학생은 'G' 로 표기하고 제적, 휴학생태, 자퇴, 수료인 경우는 모두 'H' 로 표기한 후 출신고등학교, 출신지역, 성별, 성적 등이 졸

업여부와 어떤 연관성이 있는지 분석 하였다.

본 연구의 목적은 <그림 3>에서 보는 것과 같이 학생들이 입학한 이후에 중도에 학업을 중단한 확률인 학생들의 이탈율을 효과적으로 계산 할 수 있는 Deep Learning Module을 개발하고 개발된 Module에 효과적으로 학습할 수 있는 기초 데이터를 확보하는 것에 있다.

Deep Learning 모듈을 개발한 후 개발된 모듈을 학사정보와 연계하여 학생이 입학한 경우 출신고등학교, 연령, 출신지역 등의 데이터를 모듈에 대입하면 통계 데이터의 확보 없이 이탈 확률이 높은 학생들을 선별하여 효과적인 지도학습이 가능하다.

<그림 3> 데이터 활용 방안

3.2 의사결정트리

연관 분석을 통한 데이터를 이용한 의사결정모델을 평가하기 위하여 R의 caret 패키지를 사용해서 데이터를 70% 의 train set과 30%의 test set으로 구분한 후 의사결정트리의 성능을 평가하였다. train set으로 의사결정트리 모델을 구성하고, test set을 구성된 모델이 입력하여 졸업여부를 얼마나 잘 예측하는지 확인하였다.

의사결정트리 분석은 회귀분석이나, 랜덤포레스트 등의 알고리즘에 비해 직관

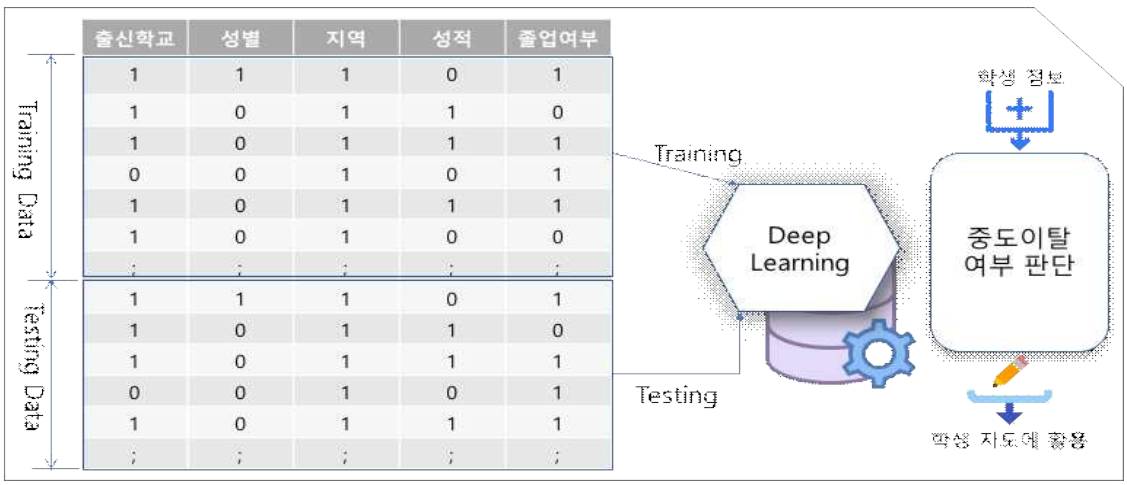
적으로 이해할 수 있으며, 설명 또한 쉽게 할수 있는 장점이 있는 지도학습 방법이다. 분석방법에는 party 패키지를 사용하여 분석하는 것이 가장 높은 정확도를 나타내었다.

3.3 Deep learning

연관 분석을 통한 데이터를 Deep Learning 모듈을 만들기 위한 데이터로 활용 하기 위하여 총 8891건의 학생 데이터를 모델을 만들기 위한 Training Data, 만들어진 모델을 평가하기에 필요한 Testing 데이터로 분리하였다. Training Data로 모델을 훈련시키고 만들어진 Deep Learning 모델이 새로 들어오는 입력에 대해 결과 예측치를 제시 하고자 한다.

연관분석을 바탕으로 한 데이터는 <그림 4>에서와 같이 각각의 분류 데이터를 바이너리 구조로 변환한 후 4,000개의 Training Data 2set, 800개의 Testing Data 로 구분하여 사용하였다.

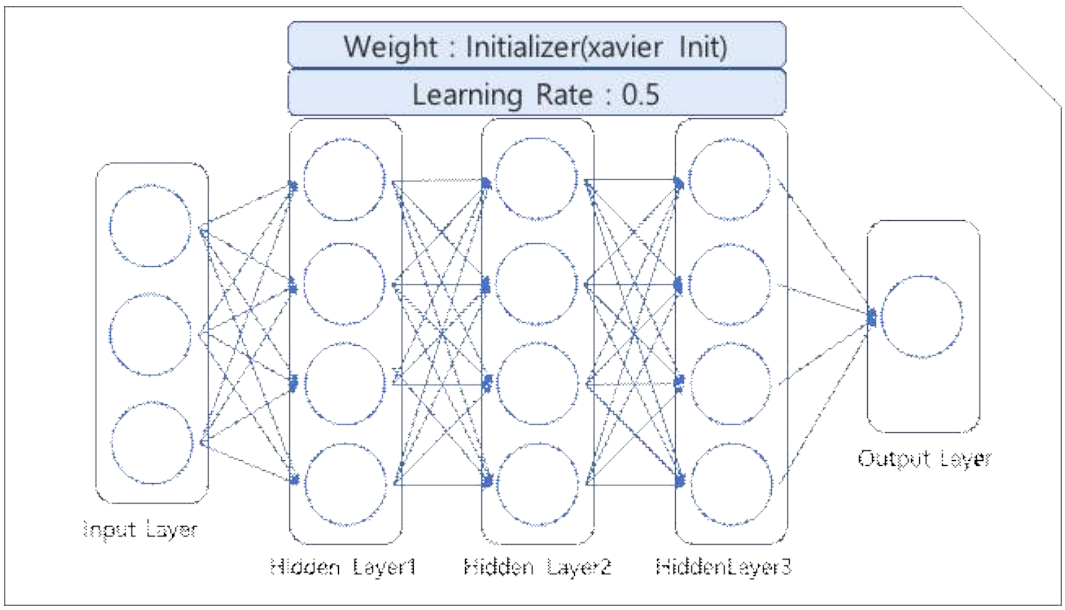
<그림 4> Training Data, Testing Data 활용



Training 및 Testing 결과로 학생들의 중도이탈여부 판단에 사용될 Deep Learning Module을 구성하고 향후에 실제 학생 데이터를 만들어진 모듈에 적용하여 학생 지도에 활용할 수 있을지에 대한 연구를 하고자 한다. Machine Learning 라이브러리는 구글에서 사용하고 있는 오픈소스 소프트웨어 라이브러

리인 TensorFlow을 사용하였다. Deep Learining 모듈의 구성은 <그림 5>에서와 같이 1개의 Input Layer, 3개의 Hidden Layer, 1개의 Output Layer로 구성하였다.

<그림 5> Neural network(NN)



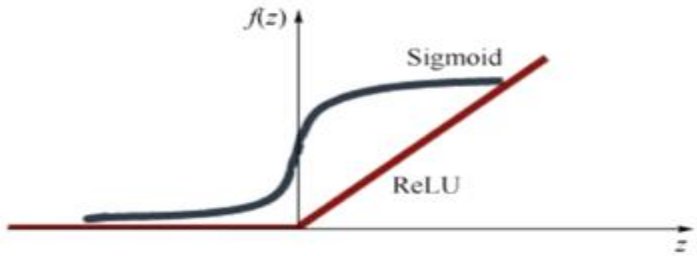
Hidden Layer를 3개보다 더 많이 구성하면 Overfitting 문제가 발생하여 정확도가 더 낮게 나타났다<그림 6>. Error loss(cost)를 위한 Learning Rate는 몇 번의 Training 결과를 반영하여 가장 큰 인식율을 나타낼 수 있는 0.5 로 결정하였다. 각각의 Layer에 입력되는 Input Data 에 각각의 Weight를 부여해야 되는데 각 Layer별 Weight 초기화 또한 중요한데 Weight에 대한 초기값을 0으로 설정하면 Deep Learning 알고리즘이 전혀 동작하지 않기 때문에 본 연구에서는 지금까지 가장 단순하면서도 가장 좋은 결과를 나타내는 Xavier 초기모듈을 사용하였다. Xavier 초기화는 입력값과 출력값 사이의 난수를 선택해서 입력값의 제곱근으로 나누는 초기화 방법이다.

<그림 6> 5 Hidden layer 구성 시 정확도

```
       [ True],
       [ True],
       ...,
       [False],
       [False],
       [False]], dtype=bool), 0.74825001]
Accuracy: 0.74825
```

각 Hidden Layer에 출력 데이터를 결정하는 함수에는 max(0,x)처럼 음수에 대해서만 0으로 처리하는 방식의 ReLu 함수를 사용하였다<그림 7>.

<그림 7> Sigmoid 와 ReLU 비교



ReLU 함수는 그림에 있는 것처럼 0보다 작을 때는 0을 사용하고, 0보다 큰 값에 대해서는 해당 값을 그대로 사용하는 방법이다. 음수에 대해서는 값이 바뀌지만, 양수에 대해서는 값을 바꾸지 않는다.

ReLU(Rectified Linear Unit)를 함수로 구현하면 다음과 같다.

0과 현재 값(x) 중에서 큰 값을 선택하는 방식이며 코드로 구현하면 max(0, x)이 된다.

Sigmoid : L2 = tf.sigmoid(tf.matmul(X,W1)+b1)

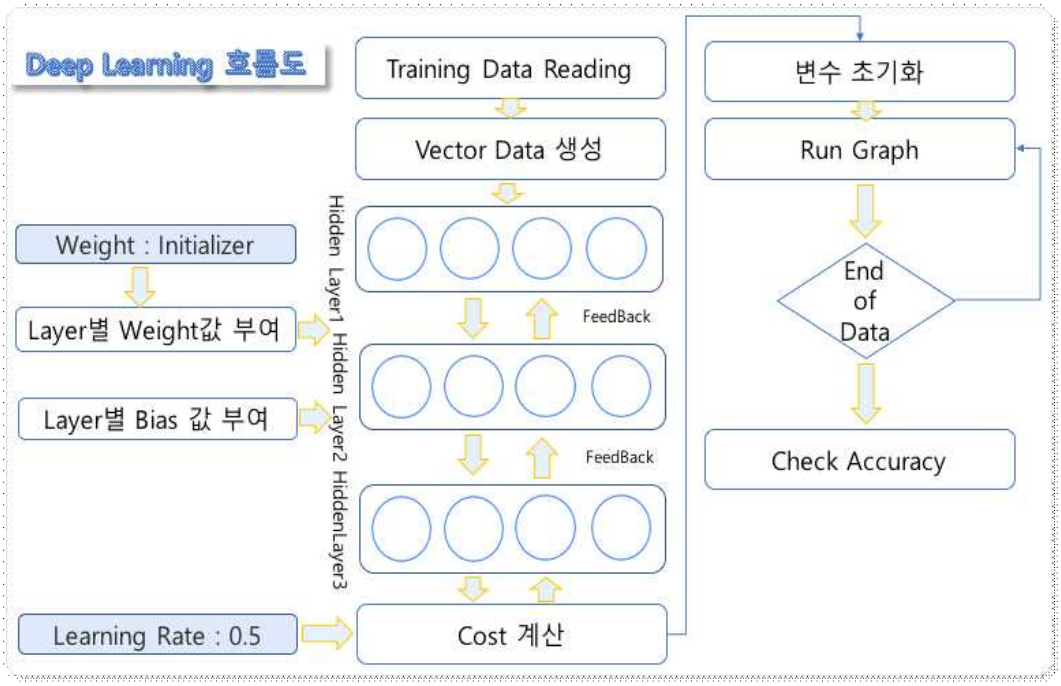
ReLU(L2 = tf.nn.relu(tf.matmul(X,W1)+b1)

설계된 Deep Learning의 흐름도를 보면 Training data를 모듈에서 활용 가능한 Vector Data로 변경한 후 각각의 Hidden Layer에 입력값으로 대입한다. 이때

각 Layer별로 Weight값을 부여한다. 출력값과 결과값을 비교한 후 경사하강법으로 손실함수를 최소화 한다. 즉 평균 제곱 오차가 최소화 되는 지점을 경사하강법으로 찾아서(cost), 각각의 Hidden Layer에 보정값으로 사용하여 학습을 진행한다.

Training Data로 학습을 진행한 후 Testing Data로 학생들의 졸업여부를 판정하는 정확도가 얼마나 되는지 측정한후 최적의 정확도가 나타날때까지 Weight 및 Learning Rate 값을 조정하여 진행 하였다<그림 8>.

<그림 8> Deep Learning 흐름도



3.4 learning Rate

learning rate 계산은 업데이트 식 $W := W - \alpha \frac{\partial}{\partial W} cost(W)$에서 α값을 의미하고 이식을 $x_1 = x_0 - t \nabla f(x_0)$ 로 표현하면 t가 learning rate를 의미한다.

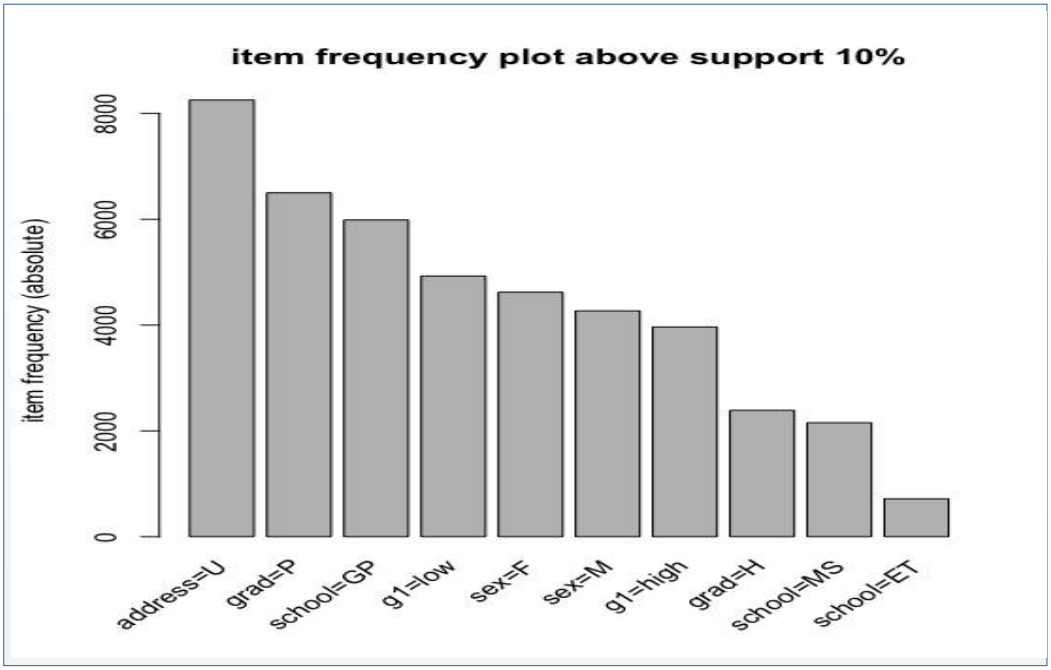
learning rate를 결정하는 방법은 정확히 나와 있지 않고, overfitting을 최소화하는 범위내에서 적당한 값을 선정하였다.

Ⅴ. 분석결과

본 연구에서는 <표 7>에서와 같이 분석에 사용된 데이터는 8891레코드에 12 컬럼의 크기를 가지는데 밀도가 0.416667정도에 불과하다. 즉 8891 * 12 * 0.4166667 = 49,789의 항목만이 분석에 포함되었다. 학업 중단율은 26.88%[3]로서 전문대학 평균보다 월등히 높아 반드시 관리해야 될 항목으로 고려 되었다.

<표 7> 데이터의 사용 빈도 및 지지도 10%인 상위 10개 item

```
transactions as itemMatrix in sparse format with
 8891 rows (elements/itemsets/transactions) and
 12 columns (items) and a density of 0.4166667
most frequent items:
address=U     grad=P school=GP     g1=low      sex=F     (Other)
     8255       6501       5988       4927       4622      14162
```



3) 2390명 / 9981명 * 100 = 26.88%

apriori 알고리즘을 사용해서 minimum support = 0.01, minimum confidence = 0.2로 하여 분석한 결과 총 577개의 rule 중에서 rule이 3개의 아이템으로 이루어져 있는 rule이 대략 206개, 4개의 아이템으로 이루어져 있는 rule이 대략 214개로 구성되어 있다.

의미 있는 지지도(support)는 0.13, 신뢰도(confidence)는 0.62, 향상도(Lift)는 1.65로서 도시지역에 거주하면서 일반고등학교를 졸업하고 여성으로 구성된 데이터가 최소 지지도와 최소 신뢰도를 넘어서는 것으로 나타나는 것으로 볼 때 전반적으로 이러한 항목들이 졸업여부에 영향을 미치는 것으로 나타났다<표 8>.

<표 8> 분석결과 summary

rulelengthdistribution (lhs + rhs):sizes **1 2 3 4 5** **9 74 206 214 74** **Min. 1st Qu. Median Mean 3rd Qu. Max.** **1.000 3.000 3.000 3.468 4.000 5.000**		
support Min. :0.01001 1st Qu.:0.03284 Median :0.09538 Mean **:0.13228** 3rd Qu.:0.18018 Max. :0.92847	confidence Min. :0.2000 1st Qu.:0.4295 Median :0.6357 Mean :0.6237 3rd Qu.:0.8218 Max. :0.9981	lift Min. : 0.4623 1st Qu.: 0.9438 Median : 1.0685 Mean : **1.6519** 3rd Qu.: 1.2255 Max. :13.1853

Right-hand side가 중도 탈락일 경우 지지도가 높은 순위로 상위 20개의 연관규칙 rule 은 <표 9>와 같다.

학생들의 출신학교가 특성화고 이면서 학기 성적이 낮은 학생들의 이탈율이 많은 것으로 나타나고 있다. Right-hand side가 졸업일 경우 지지도가 높은 상위 20개의 연관규칙 Rule은 <표 10>과 같다. <표 9>와는 반대로 학기 성적이 높은 학생들이 대부분의 지지도 상위를 차지하고 있는 것으로 볼 때 학기성적이 낮은 학생들이대학에서 학생들이 정규학기를 마치지 않고 학업을 중단하는 확률이 높은 것으로 나타났다.

<표 9> rhs 중퇴일 경우 상위 20 연관규칙

```
> inspect(sort(student_mart_tr_rule_reorder, by = "lift")[1:20])
     lhs                                      rhs      support    confidence lift
[1]  {school=MS,sex=M,gl=low}                 => {grad=H} 0.03329209 0.5441176  2.024163
[2]  {school=MS,sex=M,address=U,gl=low}       => {grad=H} 0.03284220 0.5417440  2.015333
[3]  {sex=M,address=U,gl=low}                 => {grad=H} 0.15251378 0.5323910  1.980539
[4]  {school=GP,sex=M,address=U,gl=low}       => {grad=H} 0.11697222 0.5289929  1.967898
[5]  {school=GP,sex=M,gl=low}                 => {grad=H} 0.11719717 0.5289340  1.967679
[6]  {sex=M,gl=low}                           => {grad=H} 0.16016196 0.5087531  1.892604
[7]  {school=MS,gl=low}                       => {grad=H} 0.05106287 0.4295175  1.597841
[8]  {school=MS,address=U,gl=low}             => {grad=H} 0.05038803 0.4278892  1.591784
[9]  {address=U,gl=low}                       => {grad=H} 0.20852548 0.4161616  1.548156
[10] {school=GP,sex=M,address=U}              => {grad=H} 0.14216624 0.4160632  1.547790
[11] {school=GP,sex=M}                        => {grad=H} 0.14261613 0.4157377  1.546579
[12] {sex=M,address=U}                        => {grad=H} 0.18366888 0.4134177  1.537949
[13] {school=GP,gl=low}                       => {grad=H} 0.15442582 0.4100956  1.525590
[14] {school=GP,address=U,gl=low}             => {grad=H} 0.15397593 0.4100030  1.525245
[15] {school=MS,sex=M}                        => {grad=H} 0.03902823 0.4096812  1.524048
[16] {school=MS,sex=M,address=U}              => {grad=H} 0.03846586 0.4076281  1.516411
[17] {sex=M}                                  => {grad=H} 0.19322911 0.4024362  1.497096
[18] {gl=low}                                 => {grad=H} 0.22157238 0.3998376  1.487429
[19] {school=MS,sex=F,gl=low}                 => {grad=H} 0.01777078 0.3079922  1.145757
[20] {school=MS,sex=F,address=U,gl=low}       => {grad=H} 0.01754583 0.3070866  1.142388
```

지지도 및 향상도에 따른 데이터 건 수를 확인하기 위하여 IS (Interest-Support) 측도를 계산한 결과 IS측도가 낮은 결과가 0.48정도로서 비교적 지지도와 향상도에 따른 데이터 량도 중간 수준 정도를 유지하고 있었다.

<표 10> rhs 졸업일 경우 상위 20 연관규칙

```
> inspect(sort(student_mart_tr_rule_reorder, by = "lift")[1:20])
     lhs                                      rhs      support    confidence lift
[1]  {school=MS,sex=F,gl=high}                => {grad=P} 0.08502981 0.9509434  1.300544
[2]  {school=GP,sex=F,address=U,gl=high}      => {grad=P} 0.16544821 0.9508727  1.300447
[3]  {school=MS,sex=F,address=U,gl=high}      => {grad=P} 0.08469239 0.9507576  1.300290
[4]  {school=GP,sex=F,gl=high}                => {grad=P} 0.16679789 0.9506410  1.300131
[5]  {sex=F,address=U,gl=high}                => {grad=P} 0.25598920 0.9495202  1.298598
[6]  {sex=F,gl=high}                          => {grad=P} 0.26633675 0.9494787  1.298541
[7]  {sex=F,address=C,gl=high}                => {grad=P} 0.01034754 0.9484536  1.297139
[8]  {school=ET,sex=F,gl=high}                => {grad=P} 0.01270948 0.9186992  1.256446
[9]  {school=MS,address=U,gl=high}            => {grad=P} 0.11281071 0.9184982  1.256171
[10] {school=MS,gl=high}                      => {grad=P} 0.11337307 0.9180328  1.255534
[11] {address=U,gl=high}                      => {grad=P} 0.38263412 0.8952632  1.224394
[12] {gl=high}                                => {grad=P} 0.39860533 0.8940464  1.222730
[13] {school=GP,address=U,gl=high}            => {grad=P} 0.26082555 0.8854525  1.210976
[14] {school=GP,gl=high}                      => {grad=P} 0.26285007 0.8852273  1.210668
[15] {school=ET,address=C,gl=high}            => {grad=P} 0.01338432 0.8686131  1.187946
[16] {school=ET,gl=high}                      => {grad=P} 0.01968283 0.8663366  1.184833
[17] {address=C,gl=high}                      => {grad=P} 0.01597121 0.8658537  1.184172
[18] {school=GP,sex=F,address=U}              => {grad=P} 0.28287032 0.8613014  1.177947
[19] {school=GP,sex=F}                        => {grad=P} 0.28455742 0.8611300  1.177712
[20] {sex=F,address=U}                        => {grad=P} 0.41457654 0.8562137  1.170988
```

rhs가 졸업인 경우는 중도 탈락의 경우보다 많은 데이터 량을 차지하고 있으므로 IS측도 또한 전반적으로 높은 결과를 보여주고 있다<표 11>.

남/녀의 구분에 따른 졸업여부를 확인한 결과 정규학기를 마친 여성의 비율이 79%, 남성이 44.7%로서 여성이 남성보다 정상적으로 학기를 마치는 확률이 월등히 높았다.

<표 11> 졸업/중퇴일 경우 IS(Interest-support) 측도

```
> student_rule2_df2[order(-student_rule2_df$IS), ][1:10, ]
                            rules    support confidence     lift       IS
23          {school=ET} => {grad=P} 0.06141042  0.7625698 1.042918 0.2530731
154 {school=MS,gl=high} => {grad=P} 0.11337307  0.9180328 1.255534 0.3772848
114   {sex=F,address=C} => {grad=P} 0.02969295  0.8328076 1.138977 0.1839011
263      {sex=F,gl=low} => {grad=P} 0.17793274  0.7434211 1.016729 0.4253345
17          {address=C} => {grad=P} 0.05601170  0.7830189 1.070885 0.2449124
124 {school=ET,gl=high} => {grad=P} 0.01968283  0.8663366 1.184833 0.1527117
8                    {} => {grad=P} 0.73118884  0.7311888 1.000000 0.8550958
52            {gl=high} => {grad=P} 0.39860533  0.8940464 1.222730 0.6981308
132   {school=ET,sex=F} => {grad=P} 0.03250478  0.8117978 1.110244 0.1899690
276   {sex=F,address=U} => {grad=P} 0.41457654  0.8562137 1.170988 0.6967527
```

연관분석이 완료된 데이터를 의사결정트리 모델을 이용하여 졸업여부를 예측하고 Deep Learning 모듈의 예측 정확도와 비교 분석하기 위하여 Training Data로 의사결정트리를 구성하고, Testing Data로 의사결정트리 모델을 평가한 결과 의사결정트리 모델의 예측 정확도는 75%로 측정되었다<표 12>.

Deep Learning 의 예측 정확도 80% 보다 조금 낮은 예측을 나타내었다. 비교적 직관적이고 판정기준을 명확하게 확인할 수 있는 점에서는 많은 장점이 있는 예측모델이나, 통계적인 데이터가 미리 준비되어 있어야 하는 점과, 현재의 데이터를 바로 대입하여 이탈율을 예측하지 못한다는 점에서 Deep Learning 모델 보다는 제약이 많은 것으로 나타났다.

<표 12> 의사결정트리 모델을 통한 정확도 측정

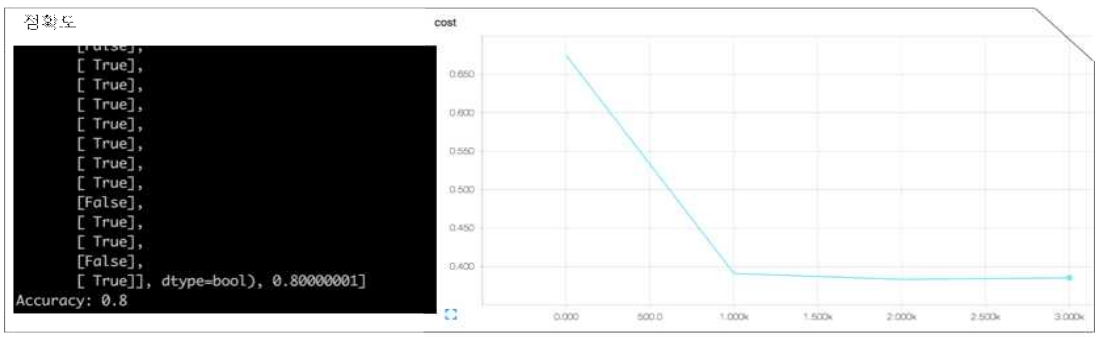
```
Confusion Matrix and Statistics
          Reference
Prediction    H    P
         H  408  358
         P  311 1590
               Accuracy : 0.7492
                 95% CI : (0.7322, 0.7655)
    No Information Rate : 0.7304
    P-Value [Acc > NIR] : 0.01490
                  Kappa : 0.3759
 Mcnemar's Test P-Value : 0.07533
            Sensitivity : 0.5675
            Specificity : 0.8162
         Pos Pred Value : 0.5326
         Neg Pred Value : 0.8364
             Prevalence : 0.2696
         Detection Rate : 0.1530
   Detection Prevalence : 0.2872
      Balanced Accuracy : 0.6918
```

연관분석이 완료된 데이터를 각각 Training Data와 Testing Data 로 분리하여 Deep Learning 모듈을 활용하여 Training 하였다.

TensorFlow Library를 활용하였으며, Weight 초기화는 xavier, Loss를 결정하는 Learning Rate는 0.5로 하여 개발된 Deep Learning Module을 Training 한 후 별도의 Testing 데이터로 Testing 한 결과 <그림9>에서와 같이 정확도가 80% 정도로서 비교적 양호한 결과를 나타내었다.

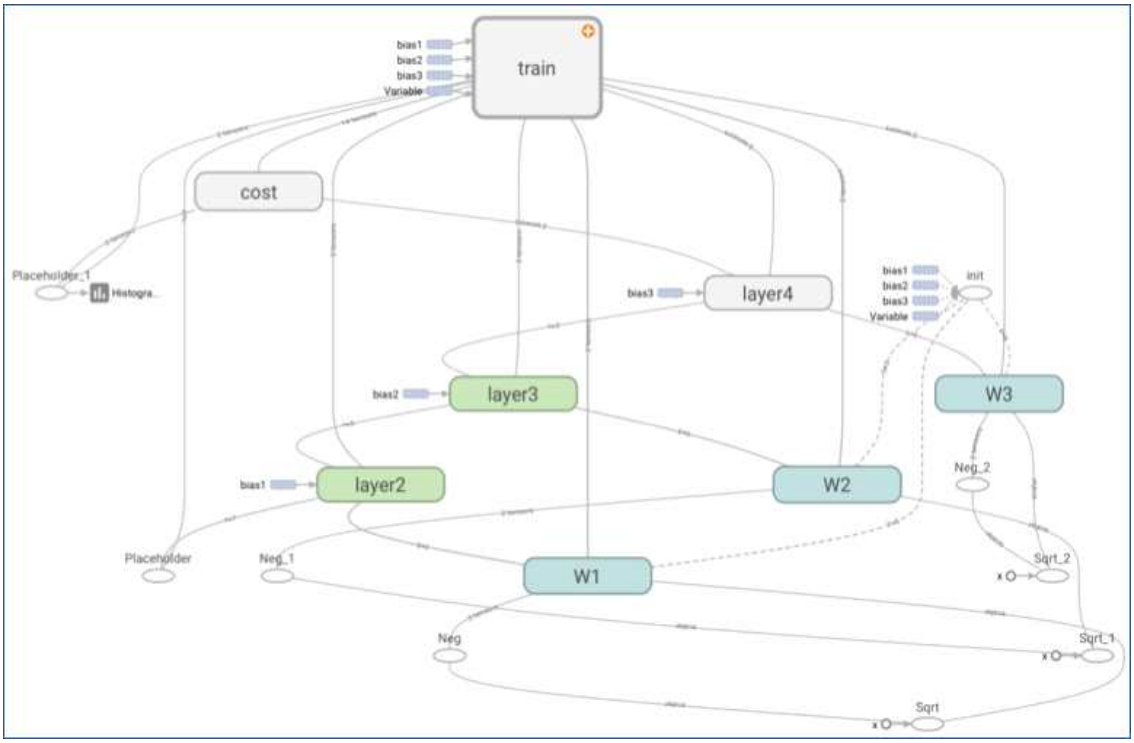
또한 Loss를 나타내는 Cost 역시 감소하는 것으로 나타났다.

<그림 9> Deep Learning 정확도 및 Cost 변화



Deep Learning 모듈, Tensorflow의 특징인 TensorBoard를 활용하여 모듈구성 Graph를 나타내고 있다<그림 10>. 각각의 Hidden Layer 별로 Weight 적용 상태, Cost 적용상태, 각 Hidden Layer별로 Training Data 의 흐름도를 나타내고 있다.

<그림 10> Deep Learning 모듈 구성 Graph



Ⅵ. 결론

본 연구에서는 일반 대학교에서 정규학기를 마치지 않고 학교를 중도에 포기하는 학생들의 연관성이 있다는 것을 알아 보았다. 특히 특성화고를 졸업하고 대학에 진학한 학생들이 학교를 중도에 포기하는 확률이 높다는 것을 확인 하였다. 이러한 결과는 특성화 고교가 학업 성취도 측면에서나 학업 지속적인 면에서 더 어려움을 겪고 있다는 것을 의미하며 이들에 대한 우선적인 관리 및 관심이 필

요할 것으로 보인다.

의사결정트리를 활용하여 졸업여부를 측정한 결과는 Deep Learning의 정확도보다 낮은 75%로 나타났다<표 13>. 의사결정트리에서 졸업여부를 결정하는 요인은 일반고를 졸업하고, 도시지역에 거주하면서 여성이면서 성적이 높은 학생들이 졸업확율이 높은 것으로 나타났다.

<표 13>에서와 같이 결과적으로 의사결정트리 보다는 개발된 Deep Learning 모듈이 더 효율적으로 학생들의 졸업여부를 평가할 수 있는 모델로 규명되었다.

<표 13> 의사결정트리와 Deep Learing 예측 정확도 비교

구분	구성	예측 정확도
의사결정트리	predict confusionMatrix	75%
Deep Learning	Hidden Layer : 3 Learning Rate : 0.5 초기화 : xavier	80%

Deep Learning 모듈의 구성은 3개의 Hidden Layer 로 구성될 경우, Learning Rate를 0.5, Weight 초기화에는 xavier초기화 모듈 사용, 각 Layer의 출력 보정에는 Relu함수를 사용한 결과 정확도는 80%정도로서 현재 까지 가장 좋은 정확도를 보여 주었다<그림 11>.

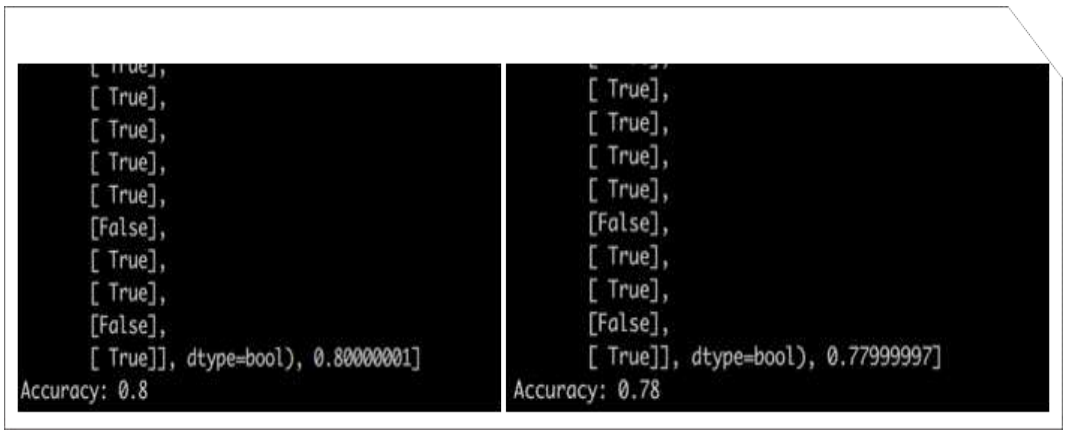
Hidden Layer를 3개 이상 구성시 overfitting의 문제로 인하여 정확도가 점점 감소하였다<그림 12>.

Deep Learning의 정확도는 80% 정도로 기대했던 만큼의 결과를 보여 주긴 하였으나, 정확도 90% 이상의 결과를 나타내기 위해서는 지역별, 대학별로 좀 더 다양한 Training 데이터를 확보하고 Deep Learning 모듈의 Layer를 좀더 넓게 적용할 필요가 있다.

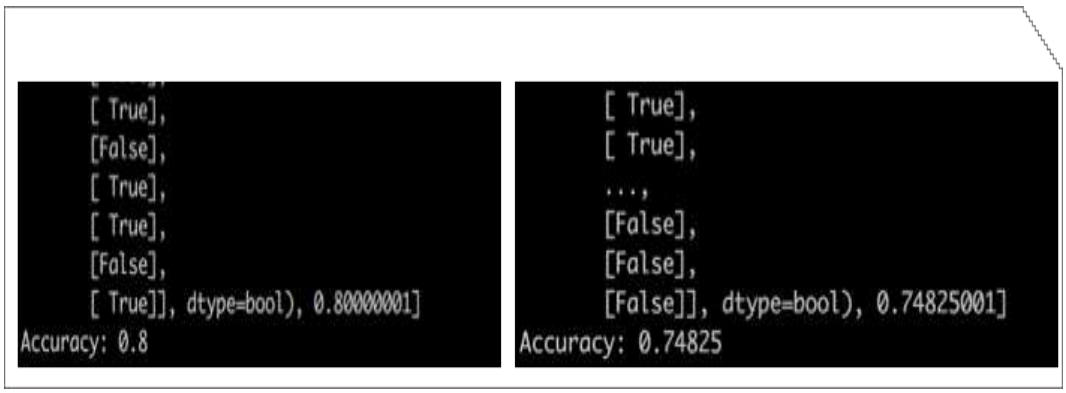
현재의 정확도만으로도 학생 이탈율 방지에 적용하여 지표의 보조 데이터로

활용할 만한 충분한 정확도가 나타나는 것으로 생각하며 향후 이들을 체계적으로 관리할 지표 및 항목을 개발하여 대학 입학에서부터 집중 관리 한다면, 불황시대에 대학이 생존할 수 있는 학생유지에 더욱 좋은 결과를 가져올 것으로 사료된다.

<그림 11> Sigmoid와 Relu 함수를 사용할때의 정확도 차이 비교



<그림 12> Hidden Layer 구성에 따른 정확도 차이 비교(Layer 3, Layer 5)



Ⅶ. 한계점 및 향후 계획

본 연구에서는 특정 대학교의 데이터만을 가지고 학생들의 기본 데이터가 학업 지속성에 어떠한 영향을 미치는가에 대한 연관성을 찾고자 하였다. 이미 존재하고 있는 학생 정보를 활용하여 이탈율에 대한 해법을 제시 하고자 하는 점, 선행 연구의 대부분이 고등학생들의 이탈율 방지에 치중하여 연구하여 왔으나 본 연구에서는 이것을 고등교육기관으로 확대하여 적용여부를 확인해 보았다는 점에서는 의미가 있으나, Deep Learning의 훈련 데이터 만을 대상으로 산정 한 결과 특정 지역의 특정 학생들만을 대상으로 한 데이터 활용, Deep Learning Training Data Sampling 수 제한, 정확도가 80%정도로서 조금 부족하다는 점 등 현재까지는 연구의 한계라고 할 수 있다.

향후 정확도를 좀 더 향상시키기 위해서 좀 더 다양한 지역의 데이터 활용이 필요(외국인 학생 데이터, 특정 지역과 기타 지역 학생 데이터, 좀 더 세분화 된 데이터 속성, 예를 들면 보호자직업 여부, 보호자 재산 여부)하다. 정확도 90% 이상의 결과를 나타내기 위해서는 지역별, 대학별로 좀 더 다양한 Training 데이터를 확보하고 Deep Learning 모듈의 Layer를 좀 더 효율적으로 적용한다면 학생 이탈율 예측에 특화된 Deep Learning 모듈로 발전시킬 수 있을 것으로 사료된다. 그러나 현재의 자료를 바탕으로 실제 업무에 적용 가능한 정확도를 가지고 있는 만큼 현재 운영되고 있는 학사정보관리에 바로 적용하여 학생 평가 시스템으로 활용할 수 있을 것이다. 또한 데이터의 속성정보 확장 등을 좀더 보완하여 전체적인 학생 이탈율 방지에 필요한 Training Data Set 과 Deep Learning Module을 필터프로그램화 할 경우 더욱 효과적인 적용이 가능할 것이다.

참 고 문 헌

권이종, 학업중단 아동과 청소년의 문제점과 해결방안, 한국아동청소년가족포럼 정책토론 자료, 2010.

김경주 외 2, 대학생의 학교이탈 관련 변인 탐색:지방 사립대학을 중심으로, 한국교육 제41권 제2호. 2014.

김성식, 대학생들의 학업중단 및 학교이동에 대한 탐색적 분석 : 대학선택요인과 대학생활 만족도의 영향, 한국교육 vol35. No.1. 2008.

김수연, 대학생의 학업지속과 중도탈락 요인분석, 한국교육, 2006.

방은주외7, 우리나라 고등학교의 정신건강이 학업 성취도와 학업 중단율에 미치는 영향, J Korean Acad Child Adolesc Psychiatry. 2016.

심현, 국립대학 학생의 중도탈락 요인에 관한 근거이론 기반 분석, 교육문화연구 제23-2호,2017.

이병환, 강대구, 중고등학교 학생들의 학교부적응 행동 요인 분석, 교육문화연구 20, 2014.

이현주,김용남, 고등학교 학업중단율 변화의 지역별, 학교유형별 현황 및 학교관련 요인 탐색, 아시아교육연구 13권 1호, pp149-185.

이희정, 도시 청소년 학교적응의 종단적 변화와 개인, 가족, 학교 및 지역사회 관련 예측 변인, 교육문화연구 21, 2015.

임연욱, 사이버대학 학습자관련 변인과 중도탈락 간의 관계 규명을 위한 실증적 연구, 한양사이버대학교.

조영재,한석근, 부모·학생 배경과 학교적응 관계 분석, 교육문제연구 21, 2015.

조완일, R의 설치 및 기본 사용법, 센소메트릭스, 2006.

Alexander Sasha Vezhnevets외 6, FeUdal Networks for Hierarchical Reinforcement Learning ,arXiv:1703.0161v2, 2017.

Chrisantha Fernando외 6, Evolution Channels Gradient Descent in Super Neural Networks, PathNet, 2017.

Convolutional Neural Networks, http://cs231n.github.io/convolutional-networks/, Marvin Minsky, 1969.

Jeffrey, V. F, An examination of first-time in college freshman attrition within the first year of attendance, Nova Southeastem Univ, Lauderdale, FL. Research and Planning. NSU-RP-R-00-25.

Jurgen Schmidhuber, Deep learning in neural networks, 2014, The Swiss AI Lab IDSIA.

Lehmann. W, "i just didn't feel like I fit in":The role of habitus in university dropout decisions, Canadian Journal of Higher Education, 37, 2007.

lembesis, A, A study of students who withdrew from college during their second, third or fourth years. 1965, Unpublished doctoral dissertation, University of Oregon.

Machine Learning, http://pythonkim.tistory.com/, Sung Kim

Nitish Srivastava 외4, A Simple Way to Prevent Neural Networks from Overfitting , Journal of Machine Learning Research 15 ,2014.

Pang-Ning Tan외 2, Introduction to Data Mining", 2006

Qingyun Sun 외 3, Convolutional Imputation of Matrix Networks, arXiv:1606.00925v2, 2017.

RFrend, R,Phthon 분석과 프로그래밍, http://rfrend.tistory.com .

TensorFlow, https://www.tensorflow.org , Google Research.

X. Glorotand Y. Bengio,Understanding the difficulty oftraining deep feedforward neural networks, in International conference onartificial intelligence and statistics, 2010.